Visualization for Text Mining in the Digital Humanities

Empowering Researchers to Use Advanced Tools for Text Mining

Julian Hocker

DIPF, Germany
Julian.Hocker@dipf.de

Abstract

In this PhD thesis, a visual interface for text analysis and text mining in the digital humanities (DH) will be developed. Text analysis is a crucial task in the DH, but advanced text mining technologies like topic modeling or clustering are difficult to use for most researchers. My work bridges this gap using visualizations. To ensure an adequate usability of visualizations for epistemological practices, the visualizations will be realized with researchers in an agile and participatory approach.

Keywords: visualization; text mining; usability; digital humanities

1 Introduction

The potentials of DH are often seen in the analysis of large corpora of texts. On the one hand there are very advanced tools for clustering and topic modelling available, but on the other hand most of the researchers in Humanities are lacking the capacities to use text mining or machine learning. Wagstaff (2012) already addresses this issue concerning machine learning in general.

In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th–15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 308–313.

How can quantitative *distant reading* (Moretti, 2016) tools become adequate epistemological tools (Ramsay & Rockwell, 2012) for humanities research in practice? How can these methods be combined with qualitative analysis? In my dissertation I follow these questions by using visualizations as a possibility to create an epistemological interface. Visualizations are regarded as a means for supporting intuitiveness and usability, in order to help people with less knowledge to be able to use these techniques. This needs to consider the concrete context in use, in order to design the adequate, new capacities. In my case, visualizations are used to enhance epistemological practices in Humanities.[1]

A research project[2] in historical educational research will be used as a case study for designing and evaluating a visual interface. The semantic research environment *Semantic CorA* (Schindler et al., 2012), which is based on *Semantic Media Wiki* and already offers tools for qualitative research, will be used as the central platform. There are only few approaches to add text analysis functions to Wikis (Witte & Sateli, 2014; Mehler et al., 2016), but all are lacking explorative visualizations and a combination with tools from close and distant reading.

My research questions are: How can a visualization interface enhance the epistemological practices in humanities? How can methods of close and distant reading be thoroughly combined? Which research capacities need to be addressed and designed to enhance humanities research? How can people with little or no knowledge about text mining be encouraged to use these methods? How can we address and create capacities for the usage of these tools in these humanities research groups?

2 State of the art

Jänicke et al. (2016) have conducted a survey of published papers for visualization in the DH. They see an emerging use of combining close and distant

1 Visualization is a very vivid field, there is some work done in DH, for example a new workshop format starting in October 2016 at the IEEE VIS conference (http://vis4dh.com/).

2 The project is called “Abiturprüfungspraxis und Abituraufsatz 1882 bis 1972” together with partners from Humboldt University Berlin and the KIT.

reading methods and introduce a taxonomy, where visualizations can be grouped based on the task they fulfil instead of the classical grouping in close or distant reading.

Researchers in DH are often not familiar with text mining technology and therefore not aware of what is possible and what is not. Because of this, visualizations should be developed together with the researchers. The developer also has to keep in mind that a visualization is not the end of the research, but is used to generate new research questions and lets the researcher dive deeper into the analysis (Jänicke, 2016). This is similar to the basic idea of explorative visualizations. Ramsay (2007) states also that the text analysis should be seen as a tool which allows analyzing the material.

There are different tools that offer text analysis methodologies for researchers in DH. But most of these tools either focus on close reading approaches (cf. Cheema et al., 2016) or do not enable users to use advanced text mining methods like clustering or topic modelling (Rockwell & Sinclair, 2016).[3] Some tools are also bound to certain corpora and therefore not reusable for other projects or corpora.

For creating visualizations with a focus on epistemological practices, the classical approach in context of usability is the visual information seeking mantra (Shneiderman, 1996). Sedlmair (2012) proposes a metric for user-centered creation of visualizations. It has three main stages, *precondition*, which is about the basic goals of the project, goals and set up, the *core stage*, which involves discovery what people want, the design and iterations of implementing code and deploying it in order to get feedback. The last stage is the *analysis* where the researchers should reflect the coding and publish the results at the end of the project. This categorization makes a lot sense because it brings agile software development to the creation of visualizations. It also means giving researches the ability to create visualizations and come to their own analysis, so researchers in the DH can be able to have a "humanistically informed theory of the making of technology", as Drucker (2012) demands. These tools have to be developed together with users from the DH in order to be useful (Borgman, 2009).

Kath et al. (2015) claim that the visualization of data is already an interpretation of data. Therefore it is necessary to make clear how researchers select the data, pre-process it using text mining algorithms and also how the visualizations are done. Some tools like *iPython* (Perez & Granger, 2007)

3 Voyant Tools, available at http://www.voyant-tools.org

and *R*[4] provide this and can be also used with little knowledge about programming, a similar tool especially for research in DH is *Voyant Notebook* (Rockwell & Sinclair, 2013). All the tools also make obvious which commands have been applied to text, but these tools have the disadvantage that users need basic coding skills and this is often not the case.

3 Research design and methods

In order to make the visualizations of an epistemological tool suitable for researchers in DH, a mix method approach will be realized in three steps. In a first step, Sedlmair's requirement analysis approach will be followed for adjusting techniques of natural language processing (NLP) to the needs of humanities researchers and for identifying the concrete problems by using these tools. Therefore there will be a participant observation and a tool analysis. Based on these results, there will be expert interviews.

The development will focus on a participative and agile approach with a close connection to researchers in digital humanities. Several visits with close contact to researches and evaluation of prototypes are planned. In a last step the developed tool will be evaluated by an expert test, which addresses the concrete new research capacities like exploration and interaction with the data, with researchers in the field of the history of education or DH.

4 Expected results

I expect my results to show an improvement for the exploration of massive text data for the non-experts through visualizations as an epistemological tool. This contains the act of interpretation of data enrichment as well as the choosing of data units to analyze. Combined with the research environment *Semantic CorA* I introduce a tool for qualitative and quantitative analysis of text corpora.

4 https://www.r-project.org

Explorative visualizations, as proposed by Jürgens et al. (2015) for patent retrieval might be a good starting point and there might be also a need to educate the DH researchers in text mining methods.

References

Borgman, C. L. (2009): The digital future is now: A call to action for the humanities. In: *Digital humanities quarterly*, 3 (4).

Cheema, M. F., Jänicke, S. Scheuermann, G. (2016): AnnotateVis: Combining Traditional Close Reading with Visual Text Analysis. In: *Workshop on Visualization for the Digital Humanities, VisWeek, 2016.*

Drucker, J. (2012): Humanistic theory and digital scholarship. In: Matthew K. Gold (Ed.): *Debates in the digital humanities* (pp. 85–95). Minneapolis, London: University of Minnesota Press.

Kath, R., Schaal, G. S. & Dumm, S. (2015): New Visual Hermeneutics. In: *Zeitschrift für germanistische Linguistik*, 43 (1), 27–51.

Jänicke, S. (2016): Valuable Research for Visualization and Digital Humanities: A Balancing Act. In: *Workshop on Visualization for the Digital Humanities, VisWeek.*

Jänicke, S., Franzini, G., Cheema, M. F. & Scheuermann, G. (2016): Visual Text Analysis in Digital Humanities. In: *Computer Graphics Forum*. doi:10.1111/cgf.12873

Jürgens, J. J., Mandl, T., Womser-Hacker, C. (2015): Visualizing Query Comparisons in Patent Retrieval Systems. In: *Proceedings of the Second International Workshop on Patent Mining and Its Applications (IPaMin 2015)*. http://ceur-ws.org/Vol-1437/ipamin2015_paper5.pdf <27.10.2016>

Mehler, A., Gleim, R., Hemati, W., Uslu, T. & Eger, S. (2016): Wikidition: Automatic lexiconization and linkification of text corpora. In: *it – Information Technology*, 58 (2), 70–79.

Moretti, F. (2013): *Distant Reading*. London, New York: Verso.

Pérez, F. & Granger, B. E. (2007): IPython: a system for interactive scientific computing. In: *Computing in Science & Engineering*, 9 (3), 21–29.

Ramsay, Stephen (2007): Algorithmic criticism. In: Ray Siemens and Susan Schreibman (Eds.): *A Companion to Digital Literary Studies.* Malden, Mass.: Blackwell. http://digitalhumanties.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&doc.view=print&chunk.id=ss1-6-7&toc.depth=1&toc.id=0 <24.10.2016>

Ramsay, S. & Rockwell, G. (2012): Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In: Gold, Matthew (Ed.): *Debates in the Digital Humanities* (pp. 75–84). Minneapolis, MN: University of Minnesota Press.

Rockwell, G. & Sinclair, S. (2016): *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, Mass.; London: The MIT Press.

Schindler, C., B. Ell, M. Rittberger (2012): Intra-linking the Research Corpus. Using Semantic MediaWiki as a lightweight Virtual Research Environment. In: Meister, Jan Christoph et al. (Eds.): *Digital Humanities 2012* (pp. 359–362). Hamburg: Hamburg University Press.

Sedlmair, M., Meyer, M. & Munzner, T. (2012): Design Study Methodology: Reflections from the Trenches and the Stacks. In: *IEEE transactions on visualization and computer graphics*, 18 (12), 2431–2440.

Shneiderman, B. (1996): The eyes have it: a task by data type taxonomy for information visualizations. In: *1996 IEEE Symposium on Visual Languages* (pp. 336–343).

Sinclair, Stéfan and Geoffrey. Rockwell (2013): Voyant Notebooks: Literate Programming, Programming Literacy. In: *Journal of Digital Humanities*, 2 (3).

Wagstaff, Kiri (2012): Machine Learning that Matters. In: *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012*. http://icml.cc/2012/papers/298.pdf

Witte, R. & Sateli, B. (2014): Adding Natural Language Processing Support to your (Semantic) MediaWiki. In: *The 9th Semantic MediaWiki Conference (SMWCon Spring 2014)*.